第二次作业

徐颖

2024-11-30

目录

Question #1:BigBangTheory. (Attached Data: BigBangTheory)

The Big Bang Theory, a situation comedy featuring Johnny Galecki, Jim Parsons, and Kaley Cuoco-Sweeting, is one of the most-watched programs on network television. The first two episodes for the

2011–2012 season premiered on September 22, 2011; the first episode attracted 14.1 million viewers and the second episode attracted 14.7 million viewers. The attached data file BigBangTheory shows the number of viewers in millions for the first 21 episodes of the 2011–2012 season (*the Big Bang theory* website, April 17, 2012).

```
### 导入数据，并对变量进行简短统计
BigBang<- read_csv("D:/xuying/data/BigBangTheory.csv")
summary(BigBang)
```

```
#>    Air Date          Viewers (millions)
#>  Length:21           Min.   :13.30
#>  Class :character    1st Qu.:14.10
#>  Mode  :character    Median :15.00
#>                      Mean   :15.04
#>                      3rd Qu.:16.00
#>                      Max.   :16.50
```

a. Compute the minimum and the maximum number of viewers.

```
print(paste("The minimum is",min(BigBang$`Viewers (millions)`,na.rm = TRUE),",the maximum is",max(
```

```
#> [1] "The minimum is 13.3 ,the maximum is 16.5"
```

b. Compute the mean, median, and mode.

```
print(paste("Mean=",mean(BigBang$`Viewers (millions)`,na.rm = TRUE),",median=",median(BigBang$`Vie
```

```
#> [1] "Mean= 15.0428571428571 ,median= 15 ,mode= 13.6"
#> [2] "Mean= 15.0428571428571 ,median= 15 ,mode= 14"
#> [3] "Mean= 15.0428571428571 ,median= 15 ,mode= 16.1"
#> [4] "Mean= 15.0428571428571 ,median= 15 ,mode= 16.2"
```

c. Compute the first and third quartiles.

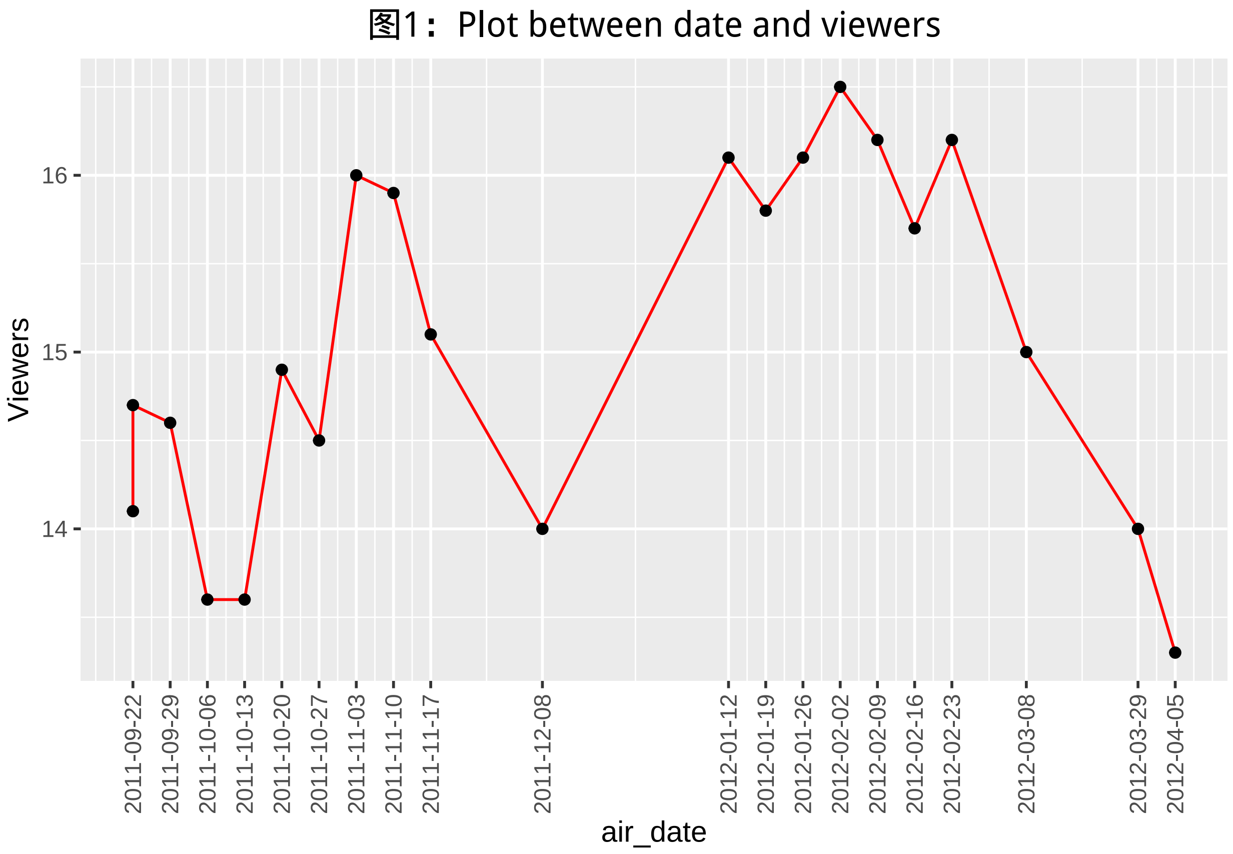
```
print(paste("q1=",quantile(BigBang$`Viewers (millions)`, 0.25, na.rm = TRUE),";q3=", quantile(BigB
```

```
#> [1] "q1= 14.1 ;q3= 16"
```

d. has viewership grown or declined over the 2011–2012 season? Discuss.

发现：2011-2012季度的收视率整体是增长的，收视率的峰值在2012年1-2月之间。

```
### 给数据集新增一列，新列为`Air Date`的日期格式
bigbang1 <- mutate(BigBang,air_date = mdy(BigBang$`Air Date`))
ggplot(data = bigbang1,mapping = aes(x=air_date,y=`Viewers (millions)`))+
  geom_line(color="red")+
  geom_point()+
  labs(title = " 图 1: Plot between date and viewers",x = "air_date", y = "Viewers") +
  scale_x_date(breaks = unique(bigbang1$air_date), date_labels = "%Y-%m-%d")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))+
  theme(plot.title = element_text(hjust = 0.5, vjust = 1))
```



Question #2: NBAPlayerPts. (Attached Data: NBAPlayerPts)

CbSSports.com developed the Total Player Rating system to rate players in the National Basketball Association (NBA) based on various offensive and defensive statistics. The attached data file NBAPlayerPts shows the average number of points scored per game (PPG) for 50 players with the highest ratings for a portion of the 2012–2013 NBA season (CbSSports.com website, February 25, 2013). Use classes starting at 10 and ending at 30 in increments of 2 for PPG in the following.

```r
### 导入数据，并对变量进行简短统计
nba<- read_csv("D:/xuying/data/NBAPlayerPts.csv")
summary(nba)
```

```
#>       Rank          Player               PPG       
#>  Min.   : 1.00   Length:50          Min.   :11.70  
#>  1st Qu.:13.25   Class :character   1st Qu.:16.30  
#>  Median :25.50   Mode  :character   Median :17.40  
#>  Mean   :25.50                      Mean   :18.29  
#>  3rd Qu.:37.75                      3rd Qu.:19.12  
#>  Max.   :50.00                      Max.   :28.80  
```

a. Show the frequency distribution.

```r
### 展示频率分布
breaks <- seq(10, 30, by = 2)
(frequency_table <- table(cut(nba$PPG, breaks = breaks)))
```

```
#> 
#> (10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30] 
#>       1       4       6      20       8       4       2       0       3       2 
```

b. Show the relative frequency distribution.

```r
### 展示相对频率分布，相对频率分布是指每个类别的频率占总频率的比例
(relative_frequency_table <- frequency_table / sum(frequency_table))
```

```
#> 
#> (10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30] 
#>    0.02    0.08    0.12    0.40    0.16    0.08    0.04    0.00    0.06    0.04 
```

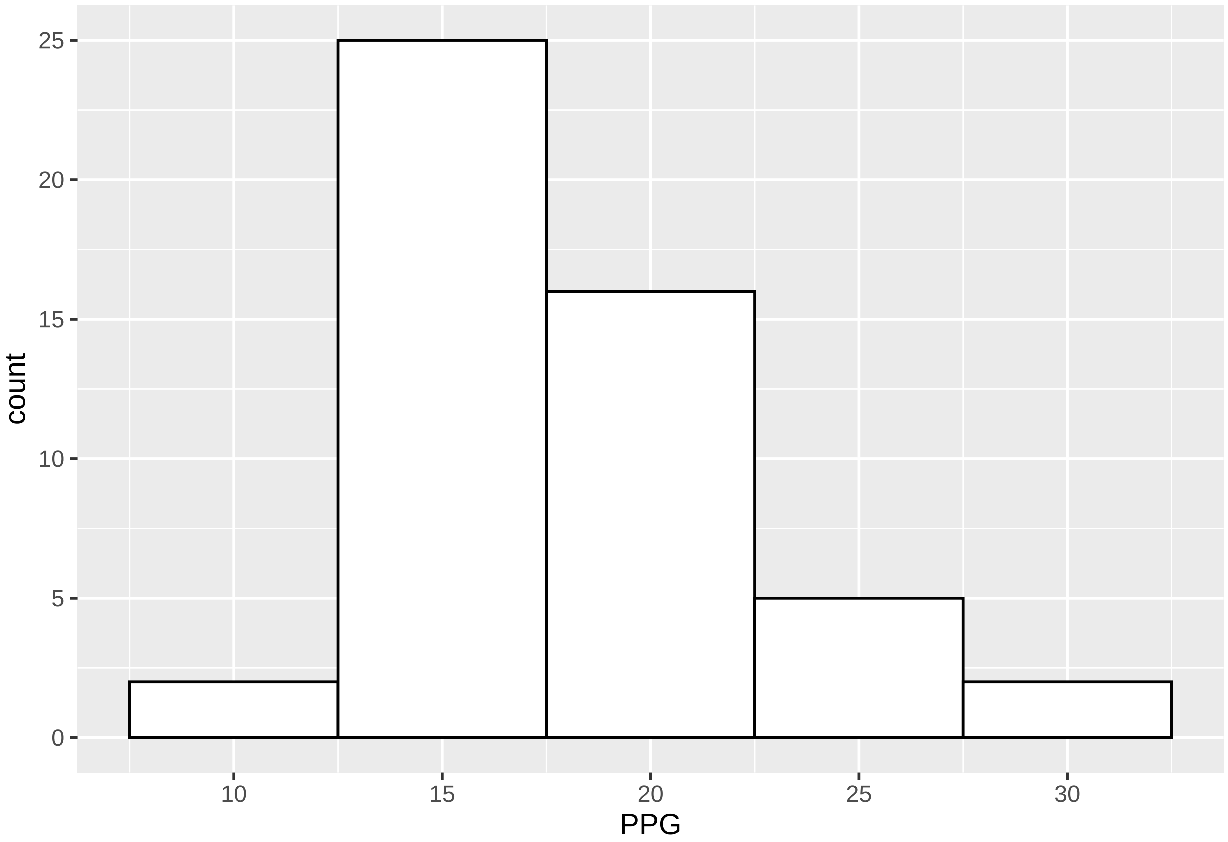
c. Show the cumulative percent frequency distribution.

```
### 展示累积分布百分比，累积分布百分比是指每个类别的频率占总频率的累积比例
cumulative_frequency <- cumsum(frequency_table)
(cumulative_percent_frequency <- cumulative_frequency / sum(frequency_table))
```

```
#> (10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30]
#>    0.02    0.10    0.22    0.62    0.78    0.86    0.90    0.90    0.96    1.00
```

d. Develop a histogram for the average number of points scored per game.

```
### 为每场比赛的平均分创建一个直方图，宽度设置为 5
ggplot(data = nba,mapping = aes(x=PPG))+
geom_histogram(binwidth = 5,fill="white",color="black")
```



e. Do the data appear to be skewed? Explain.

从上面的直方图可以看出，数据右边尾巴比较长，应该是右偏。经核算，数据的偏度值为 1.124025，且偏度值大于 0，因此证实数据有偏度，且为右偏。

```r
### 数据是否存在偏度，计算数据的偏度值
(skewness_value <- skewness(nba$PPG, na.rm = TRUE))
```

```
#> [1] 1.124025
```

f. What percentage of the players averaged at least 20 points per game?

```r
### 计算平均得分不低于 20 的球员占比
cat(sprintf("%.0f%%\n",sum(nba$PPG>=20)/max(row_number(nba))*100))
```

```
#> 22%
```

Question #3: A researcher reports survey results by stating that the standard error of the mean is 20. The population standard deviation is 500.

a. How large was the sample used in this survey?

```r
### 标准误差 = 标准差/样本量开平方
(sample_size <- (500/20)^2)
```

```
#> [1] 625
```

b. What is the probability that the point estimate was within ±25 of the population mean?

```r
### pnorm(upper_bound, mean = mu, sd = se)
### 均值 mu 如果未知，可以默认设置为 0
round(pnorm(25,0,20)-pnorm(-25,0,20),2)
```

```
#> [1] 0.79
```

Question #4: Young Professional Magazine (Attached Data:Professional)

Young Professional magazine was developed for a target audience of recent college graduates who are in their first 10 years in a business/professional career. In its two years of publication, the magazine has been fairly successful. Now the publisher is interested in expanding the magazine's advertising base. Potential advertisers continually ask about the demographics and interests of subscribers to *young Professionals*. To collect this information, the magazine commissioned a survey to develop a profile of its subscribers. The survey results will be used to help the magazine choose articles of interest and provide advertisers with a profile of subscribers. As a new employee of the magazine, you have been asked to help analyze the survey results.

Some of the survey questions follow:

1. What is your age?
2. Are you: Male______________ Female_________________
3. Do you plan to make any real estate purchases in the next two years? Yes_________ No_________
4. What is the approximate total value of financial investments, exclusive of your home, owned by you or members of your household?
5. How many stock/bond/mutual fund transactions have you made in the past year?
6. Do you have broadband access to the Internet at home? Yes_________ No_________
7. Please indicate your total household income last year. _________________
8. Do you have children? Yes_________ No_________ The file entitled Professional contains the responses to these questions.

Managerial Report:

Prepare a managerial report summarizing the results of the survey. In addition to statistical summaries, discuss how the magazine might use these results to attract advertisers. You might also comment on how the survey results could be used by the magazine's editors to identify topics that would be of interest to readers. Your report should address the following issues, but do not limit your analysis to just these areas.

```r
### 导入数据，并对变量进行简短统计
professional<- read_csv("D:/xuying/data/Professional.csv")
### summary(professional)
```

```r
professional2 <- professional %>%
  select(Age:`Have Children?`)
```

```
professional_skim <- as.data.frame(skim(professional2))
professional_select <- professional_skim %>%
  select(skim_type,skim_variable,n_missing,complete_rate,character.n_unique,numeric.mean:numeric.h
professional_select
```

a. Develop appropriate descriptive statistics to summarize the data.

```
#>   skim_type            skim_variable n_missing complete_rate character.n_unique
#> 1 character                   Gender         0             1                  2
#> 2 character   Real Estate Purchases?         0             1                  2
#> 3 character        Broadband Access?         0             1                  2
#> 4 character           Have Children?         0             1                  2
#> 5   numeric                      Age         0             1                 NA
#> 6   numeric Value of Investments ($)         0             1                 NA
#> 7   numeric   Number of Transactions         0             1                 NA
#> 8   numeric     Household Income ($)         0             1                 NA
#>   numeric.mean   numeric.sd numeric.p0 numeric.p25 numeric.p50 numeric.p75
#> 1           NA           NA         NA          NA          NA          NA
#> 2           NA           NA         NA          NA          NA          NA
#> 3           NA           NA         NA          NA          NA          NA
#> 4           NA           NA         NA          NA          NA          NA
#> 5    30.112195     4.024023         19          28          30          33
#> 6 28538.292683 15810.830741          0       18300       24800       34275
#> 7     5.973171     3.100873          0           4           6           7
#> 8 74459.512195 34818.210672      16200       51625       66050       88775
#>   numeric.p100 numeric.hist
#> 1           NA         <NA>
#> 2           NA         <NA>
#> 3           NA         <NA>
#> 4           NA         <NA>
#> 5           42
#> 6       133400
#> 7           21
#> 8       322500
```

```
### 不确定样本是否完全符合正态分布或样本量小于 30 时，用 t 检验
### t.test(professional2$Age)
```

```
mean_age <- mean(professional2$Age)
sd_age <- sd(professional2$Age)
n_age <- length(professional2$Age)
se_age <- sd_age/sqrt(n_age)
alpha <- 0.05
t_value <- qt(1-alpha/2,df=n_age-1)
ci_lower <- mean_age-t_value*se_age
ci_upper <- mean_age+t_value*se_age
print(paste("95% confidence intervals for the mean age from ",round(ci_lower,1),"to",round(ci_uppe
```

b. Develop 95% confidence intervals for the mean age and household income of subscribers.

```
#> [1] "95% confidence intervals for the mean age from  29.7 to 30.5"
```

```
### 不确定样本是否完全符合正态分布或样本量小于 30 时，用 t 检验
### t.test(professional2$`Household Income ($)`)
mean_income <- mean(professional2$`Household Income ($)`)
sd_income <- sd(professional2$`Household Income ($)`)
n_income <- length(professional2$`Household Income ($)`)
se_income <- sd_income/sqrt(n_income)
ci_lower_income <- mean_income-t_value*se_income
ci_upper_income <- mean_income+t_value*se_income
print(paste("95% confidence intervals for the household income from ",round(ci_lower_income,0),"to
```

```
#> [1] "95% confidence intervals for the household income from  71079 to 77840"
```

```


### 可以使用 prop.test() 函数来进行二项分布的比例检验
k_broadhand <- sum(professional2$`Broadband Access?`=="Yes")
n_broadhand <- length(professional2$`Broadband Access?`)
prop_broadhand <- prop.test(x=k_broadhand,n=n_broadhand,conf.level = 0.95)
round(prop_broadhand$conf.int[1],3)
```

c. Develop 95% confidence intervals for the proportion of subscribers who have broadband access at home and the proportion of subscribers who have children.

```
#> [1] 0.575
```

```r
round(prop_broadhand$conf.int[2],3)
```

```
#> [1] 0.671
```

95% confidence intervals for the proportion of subscribers who have broadband access at home from 0.575 to 0.671.

```r
### 可以使用 prop.test() 函数来进行二项分布的比例检验
k_child <- sum(professional2$`Have Children?`=="Yes")
n_child <- length(professional2$`Have Children?`)
prop_child <- prop.test(x=k_child,n=n_child,conf.level = 0.95)
round(prop_child$conf.int[1],3)
```

```
#> [1] 0.485
```

```r
round(prop_child$conf.int[2],3)
```

```
#> [1] 0.583
```

95% confidence intervals for the proportion of subscribers who have broadband access at home from 0.485 to 0.583.

```r


round(mean(professional2$`Value of Investments ($)`),0)
```

d. Would *Young Professional* be a good advertising outlet for online brokers? Justify your conclusion with statistical data.

```
#> [1] 28538
```

```r
round(mean(professional2$`Number of Transactions`),0)
```

```
#> [1] 6
```

Most of the subscribers have financial investments exclusive of their home(the mean investments is $ 28538) and they often transact stock/bond/mutual fund (the mean number of transactions is 6 in the last year). So, Young Professional should be a good advertising outlet for online brokers.

```
round(sum(professional2$`Have Children?`=="Yes")/length(professional2$`Have Children?`)*100,1)
```

e. Would this magazine be a good place to advertise for companies selling educational software and computer games for young children?

```
#> [1] 53.4
```

```
round(sum(professional2$`Broadband Access?`=="Yes")/length(professional2$`Broadband Access?`)*100,
```

```
#> [1] 62.4
```

The proportion of the subscribers who have children is 53.4% and the proportion of the subscribers who have broadband access at home is 62.4% . So, this magazine is a good place to advertise for companies selling educational software and computer games for young children.

```
round(mean(professional2$Age),1)
```

f. Comment on the types of articles you believe would be of interest to readers of *Young Professional*.

```
#> [1] 30.1
```

```
round(mean(professional2$`Value of Investments ($)`),0)
```

```
#> [1] 28538
```

```
round(mean(professional2$`Number of Transactions`),0)
```

```
#> [1] 6
```

```
round(sum(professional2$`Have Children?`=="Yes")/length(professional2$`Have Children?`)*100,1)
```

```
#> [1] 53.4
```

As the mean age of the subscribers is 30.1 , so the Career Planning articles maybe a good choice;

As the mean financial investments of the subscribers is $ 28538 and the frequency of stock/bond/mutual fund transactions is 6 , so the Investments articles maybe an another good choice;

As the proportion of the subscribers who have children is 53.4% , so the Parenting articles maybe a good choice too.

Question #5: Quality Associate, Inc. (Attached Data: Quality)

Quality associates, inc., a consulting firm, advises its clients about sampling and statistical procedures that can be used to control their manufacturing processes. in one particular application, a client gave Quality associates a sample of 800 observations taken during a time in which that client's process was operating satisfactorily. the sample standard deviation for these data was .21; hence, with so much data, the population standard deviation was assumed to be .21. Quality associates then suggested that random samples of size 30 be taken periodically to monitor the process on an ongoing basis. by analyzing the new samples, the client could quickly learn whether the process was operating satisfactorily. when the process was not operating satisfactorily, corrective action could be taken to eliminate the problem. the design specification indicated the mean for the process should be 12. the hypothesis test suggested by Quality associates follows.

$$H_0 : \mu = 12 H_1 : \mu \neq 12$$

Corrective action will be taken any time H_0 is rejected. Data are available in the data set Quality.

Managerial Report

```
### 导入数据，并对变量进行简短统计
quality<- read_csv("D:/xuying/data/Quality.csv")
summary(quality)
```

```
#>     Sample 1         Sample 2         Sample 3         Sample 4
#>  Min.   :11.52   Min.   :11.59   Min.   :11.36   Min.   :11.64
#>  1st Qu.:11.81   1st Qu.:11.88   1st Qu.:11.75   1st Qu.:11.98
#>  Median :11.96   Median :12.03   Median :11.92   Median :12.08
#>  Mean   :11.96   Mean   :12.03   Mean   :11.89   Mean   :12.08
#>  3rd Qu.:12.14   3rd Qu.:12.21   3rd Qu.:12.00   3rd Qu.:12.23
#>  Max.   :12.32   Max.   :12.39   Max.   :12.22   Max.   :12.47
```

```


### 用 t 检验的 p 值或者置信区间来判断是否拒绝原假设 H0
# 假设检验的结果存储
results <- data.frame(Sample = character(), P_Value = numeric(), Decision = character(), stringsAs
# 对每个样本进行假设检验
for (sample_name in names(quality)) {
  sample_data <- quality[[sample_name]]
  # 进行 t 检验，均值为 12（根据实际情况更改均值）
  t_test_result <- t.test(sample_data, mu = 12)
  # 提取 p 值
  p_value <- t_test_result$p.value
  # 决策
  if (p_value < 0.01) {
    decision <- " 拒绝零假设"
  } else {
    decision <- " 不拒绝零假设"
  }
  # 存储结果
results <- rbind(results, data.frame(Sample = sample_name, P_Value = p_value,  Decision = decision
}
# 打印结果
print(results)
```

a. Conduct a hypothesis test for each sample at the .01 level of significance and determine what action, if any, should be taken.Provide the p-value for each test.

```
#>     Sample     P_Value     Decision
#> 1 Sample 1 0.312729582 不拒绝零假设
#> 2 Sample 2 0.481820940 不拒绝零假设
#> 3 Sample 3 0.006468822   拒绝零假设
#> 4 Sample 4 0.039058947 不拒绝零假设
```

```


### 卡方检验可以用来检验样本方差、标准差是否显著偏离假设的总体方差或标准差
# 总体标准差
sigma <- 0.21
# 计算每个样本的标准差和进行卡方检验
results <- data.frame(Sample = character(), Sample_SD = numeric(), Chi_Square = numeric(), P_Value

for (i in 1:length(quality)) {
    sample_data <- quality[[i]]
    sample_sd <- sd(sample_data)  # 计算样本标准差
    n <- length(sample_data)  # 样本大小

    # 计算卡方统计量
    chi_square <- (n - 1) * (sample_sd^2) / (sigma^2)
    # 计算 p 值
    p_value <- 1 - pchisq(chi_square, df = n - 1)  # 右尾检验

    # 决策
    decision <- ifelse(p_value < 0.01, "Reject H0", "Fail to Reject H0")

    # 存储结果
    results <- rbind(results, data.frame(Sample = paste("Sample", i), Sample_SD = sample_sd, Chi_S
}

# 查看结果
print(results)
```

b. compute the standard deviation for each of the four samples. does the assumption of .21 for the population standard deviation appear reasonable?

```
#>     Sample Sample_SD Chi_Square   P_Value          Decision
#> 1 Sample 1 0.2203560   31.93076 0.3229206 Fail to Reject H0
#> 2 Sample 2 0.2203560   31.93076 0.3229206 Fail to Reject H0
#> 3 Sample 3 0.2071706   28.22381 0.5059716 Fail to Reject H0
#> 4 Sample 4 0.2061090   27.93530 0.5213746 Fail to Reject H0
```

```
### 用 t 检验的 p 值或者置信区间来判断是否拒绝原假设 H0
t_test_result_5c <- t.test(quality,conf.level = 0.99)
t_test_result_5c$conf.int
```

c. compute limits for the sample mean $\overline{x}$ around $\mu = 12$ such that, as long as a new sample mean is within those limits, the process will be considered to be operating satisfactorily. if $\overline{x}$ exceeds the upper limit or if $\overline{x}$ is below the lower limit, corrective action will be taken. these limits are referred to as upper and lower control limits for quality control purposes.

```
#> [1] 11.93610 12.04273
#> attr(,"conf.level")
#> [1] 0.99
```

```
t_test_result_5d <- t.test(quality,conf.level = 0.95)
t_test_result_5d$conf.int
```

d. discuss the implications of changing the level of significance to a larger value. what mistake or error could increase if the level of significance is increased?

```
#> [1] 11.94909 12.02975
#> attr(,"conf.level")
#> [1] 0.95
```

当显著性水平为 1% 时，对应的置信区间为 11.936-12.0427，当显著性水平增加至 5% 时，对应的置信区间为 11.949-12.0297，显著性水平越大，对应的置信区间越小，会增加第一类错误的风险，即：增加了弃真的风险。

Question #6: Vacation occupancy rates were expected to be up during March 2008 in Myrtle Beach, South Carolina (*the sun news*, February 29, 2008). Data in the file Occupancy (Attached file Occupancy) will allow you to replicate the findings presented in the newspaper. The data show units rented and not rented for a random sample of vacation properties during the first week of March 2007 and March 2008.

```
### 导入数据，并对变量进行简短统计，导入时删除异常的第一行
occupancy<- read_csv("D:/xuying/data/Occupancy.csv",skip = 1)
```

a. Estimate the proportion of units rented during the first week of March 2007 and the first week of March 2008.

```
rented_prop_2007 <- sum(occupancy$`March 2007`=="Yes")/length(occupancy$`March 2007`)
rented_prop_2008 <- sum(occupancy$`March 2008`=="Yes",na.rm = TRUE)/length(na.omit(occupancy$`Marc
rented_prop_2007
```

```
#> [1] 0.35
```

```
rented_prop_2008
```

```
#> [1] 0.4666667
```

The proportion of units rented during the first week of March 2007 is 0.35 , the proportion of units rented during the first week of March 2008 is 0.47.

b. Provide a 95% confidence interval for the difference in proportions.

```
### 用于进行比例的假设检验。可以比较一个或两个比例的差异，并提供检验的统计量、p 值和置信区间
x=c(sum(occupancy$`March 2007`=="Yes"),sum(occupancy$`March 2008`=="Yes",na.rm = TRUE))
n=c(length(occupancy$`March 2007`),length(na.omit(occupancy$`March 2008`)))
occ_prop <- prop.test(x,n,conf.level = 0.95)
print(paste("95% confidence intervals for the difference in proportions  from ",round(occ_prop$con
```

```
#> [1] "95% confidence intervals for the difference in proportions  from  -0.23 to -0.01"
```

c. On the basis of your findings, does it appear March rental rates for 2008 will be up from those a year earlier?

由于这个置信区间的下限是 -0.23，且上限是 -0.01，且整个区间都小于 0，这意味着在 95% 的置信水平下，我们可以认为 2008 年的出租比例显著低于 2007 年的出租比例。如果出租比例下降，通常可以推测出租金的上涨压力可能较小，甚至可能出现下降的趋势。这是因为出租比例下降可能意味着市场供给过剩，或者需求减弱，从而对租金产生下行压力。

Question #7: Air Force Training Program (data file: Training)

An air force introductory course in electronics uses a personalized system of instruction whereby each student views a videotaped lecture and then is given a programmed instruc-tion text. the students work independently with the text until they have completed the training and passed a test. Of concern is the varying pace at which the students complete this portion of their training program. Some students are able to cover the programmed instruction text relatively quickly, whereas other students work much longer with the text and require additional time to complete the course. The fast students wait until the slow students complete the introductory course before the entire group proceeds together with other aspects of their training.

A proposed alternative system involves use of computer-assisted instruction. In this method, all students view the same videotaped lecture and then each is assigned to a computer terminal for further instruction. The computer guides the student, working independently, through the self-training portion of the course.

To compare the proposed and current methods of instruction, an entering class of 122 students was assigned randomly to one of the two methods. one group of 61 students used the current programmed-text method and the other group of 61 students used the proposed computer-assisted method. The time in hours was recorded for each student in the study. Data are provided in the data set training (see Attached file).

Managerial Report

```
### 导入数据，并对变量进行简短统计
training<- read_csv("D:/xuying/data/Training.csv")
summary(training)
```

```
#>     Current         Proposed    
#>  Min.   :65.00   Min.   :69.00  
#>  1st Qu.:72.00   1st Qu.:74.00  
#>  Median :76.00   Median :76.00  
#>  Mean   :75.07   Mean   :75.43  
#>  3rd Qu.:78.00   3rd Qu.:77.00  
#>  Max.   :84.00   Max.   :82.00  
```

```
skimr::skim(training)
```

a. use appropriate descriptive statistics to summarize the training time data for each method. what similarities or differences do you observe from the sample data?

表 1: Data summary

Name	training
Number of rows	61
Number of columns	2
Column type frequency:	
numeric	2
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Current	0	1	75.07	3.94	65	72	76	78	84	
Proposed	0	1	75.43	2.51	69	74	76	77	82	

通过以上描述性统计数据可以看出，两种方法的用时中位数是相同的，用时平均值也是接近的，但是建议方案的用时更集中，尤其是右尾部分耗时长的占比明显减少。

b. Comment on any difference between the population means for the two methods. Discuss your findings.

```
### 当有两个独立的样本时，可以使用独立样本 t 检验。它用于检验两个样本均值是否存在显著差异。
t_train <- t.test(training$Current,training$Proposed)
alpha <- 0.05
if (t_train$p.value < alpha) {
    print(" 拒绝零假设：两种方法的均值存在显著差异。")
} else {
    print(" 未拒绝零假设：两种方法的均值没有显著差异。")
}
```

```
#> [1] "未拒绝零假设：两种方法的均值没有显著差异。"
```

```


### F 检验用于比较两个样本的方差是否相等。
map(training,sd)
```

c. compute the standard deviation and variance for each training method. conduct a hypothesis test about the equality of population variances for the two training methods. Discuss your findings.

```
#> $Current
#> [1] 3.944907
#>
#> $Proposed
#> [1] 2.506385
```

```
map(training,var)
```

```
#> $Current
#> [1] 15.5623
#>
#> $Proposed
#> [1] 6.281967
```

```
f_train <- var.test(training$Current,training$Proposed)
alpha <- 0.05
if (f_train$p.value < alpha) {
    print(" 拒绝零假设：两种方法的方差存在显著差异。")
} else {
    print(" 未拒绝零假设：两种方法的方差没有显著差异。")
}
```

```
#> [1] "拒绝零假设：两种方法的方差存在显著差异。"
```

d. what conclusion can you reach about any differences between the two methods? what is your recommendation? explain. 建议用建议方案，原因如下：两种方法的用时平均值是接近的 (平均值的 p.value=0.5481>0.05)，但是两种方法的标准差是不一样的（标准差的 p.value=0.000578<0.05），建议方案的方差为 6.28 明显比当前方案的方差 15.56 小，说明建议方案的用时更集中，尤其是右尾部分耗时长的占比明显减少，可以有效缩短快学生的等待时间。

e. can you suggest other data or testing that might be desirable before making a final decision on the training program to be used in the future? 建议对比一下学习成绩，如果学习成绩更好且用时更集中，则可以切换。

Question #8: The Toyota Camry is one of the best-selling cars in North America. The cost of a previously owned Camry depends upon many factors, including the model year, mileage, and condition. To investigate the relationship between the car's mileage and the sales price for a 2007 model year Camry, Attached data file Camry show the mileage and sale price for 19 sales (Pricehub website, February 24, 2012).

```
### 导入数据，并对变量进行简短统计
camry<- read_csv("D:/xuying/data/Camry.csv")
summary(camry)
```

```
#>  Miles (1000s)     Price ($1000s)
#>  Min.   : 22.00   Min.   : 8.30
#>  1st Qu.: 44.50   1st Qu.:11.35
#>  Median : 68.00   Median :12.50
#>  Mean   : 66.74   Mean   :12.55
#>  3rd Qu.: 89.00   3rd Qu.:13.40
#>  Max.   :110.00   Max.   :16.20
```

```
skimr::skim(camry)
```

表 3: Data summary

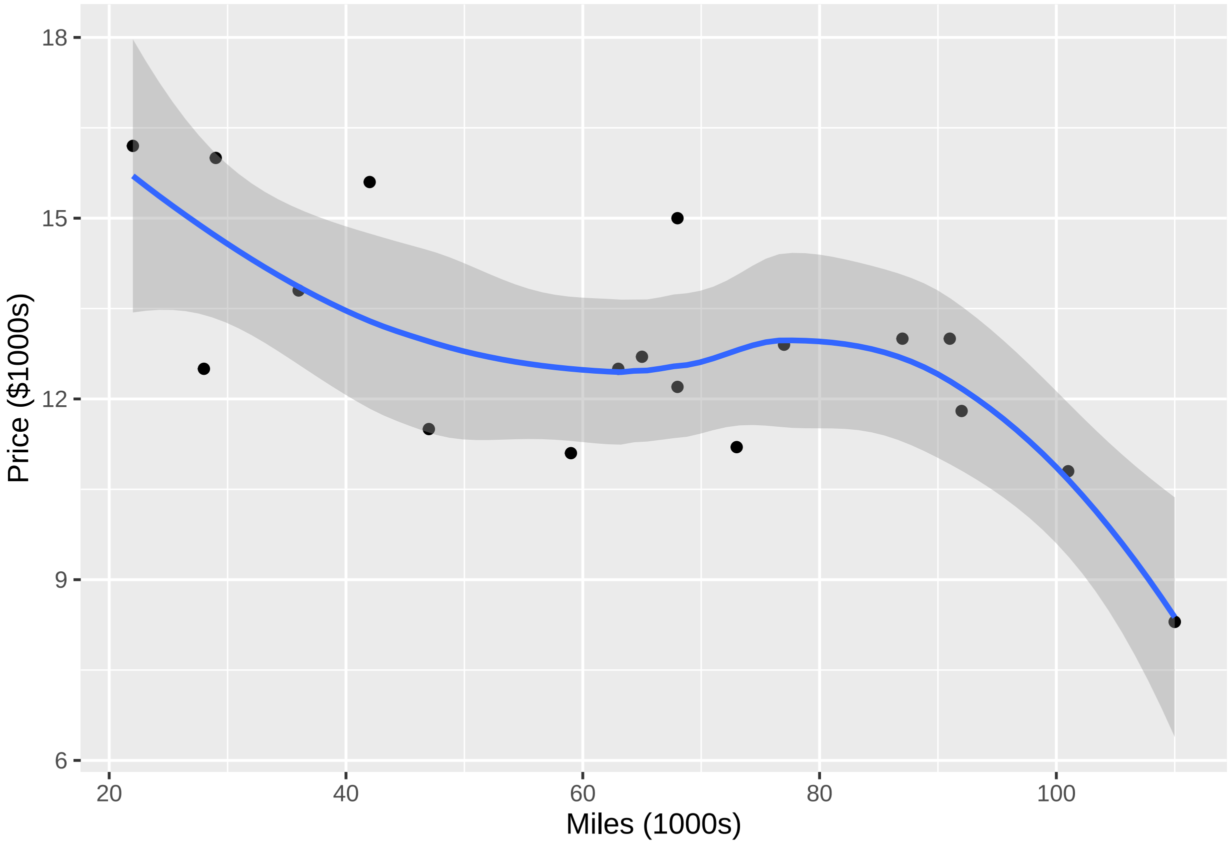
Name	camry
Number of rows	19
Number of columns	2
Column type frequency:	
numeric	2
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Miles (1000s)	0	1	66.74	27.54	22.0	44.50	68.0	89.0	110.0	
Price ($1000s)	0	1	12.55	2.21	8.3	11.35	12.5	13.4	16.2	

a. Develop a scatter diagram with the car mileage on the horizontal axis and the price on the vertical axis.

```
ggplot(data = camry,mapping = aes(x=`Miles (1000s)`,y=`Price ($1000s)`))+
  geom_point()+
  geom_smooth()
```



b. what does the scatter diagram developed in part (a) indicate about the relationship between the two variables?

给 a 部分的散点图加上曲线图之后，发现两个变量是负相关的。当历程大于 9w 公里之后，价格的降幅明显增大。

c. Develop the estimated regression equation that could be used to

predict the price ($1000s) given the miles (1000s).

```
### 使用 lm() 函数来拟合线性回归模型
lm_camry <- lm(`Price ($1000s)`~`Miles (1000s)`,camry)
summary(lm_camry)
```

```
#>
#> Call:
#> lm(formula = `Price ($1000s)` ~ `Miles (1000s)`, data = camry)
#>
```

```
#> Residuals:
#>      Min       1Q   Median       3Q      Max 
#> -2.32408 -1.34194  0.05055  1.12898  2.52687 
#> 
#> Coefficients:
#>                 Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)     16.46976    0.94876  17.359 2.99e-12 ***
#> `Miles (1000s)` -0.05877    0.01319  -4.455 0.000348 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> Residual standard error: 1.541 on 17 degrees of freedom
#> Multiple R-squared:  0.5387, Adjusted R-squared:  0.5115 
#> F-statistic: 19.85 on 1 and 17 DF,  p-value: 0.0003475
```

```r
print(paste("回归方程：Price=16.46976-0.05877*Miles"))
```

```
#> [1] "回归方程：Price=16.46976-0.05877*Miles"
```

d. Test for a significant relationship at the .05 level of significance.

c 部分的回归方程中的 p 值为 0.000348 明显小于 0.05，说明 Miles 对 Price 有显著影响。

e. Did the estimated regression equation provide a good fit? Explain.

```r
### 判断所估计的回归方程是否能够有效地解释因变量的变异，以及模型的预测能力如何
### 拟合优度的指标：
### R²（决定系数）：R² 表示自变量对因变量变异的解释比例。值越接 1，说明模型拟合越好。
### 调整后的 R²：在比较包含不同数量自变量的模型时，调整后的 R² 更为合适，因为它考虑了模型复杂度。
R_squared <- 0.5387
Adjusted_R_squared <- 0.5115
```

由于 R^2=0.5387 不接近 1，说明估计的里程对价格的回归方程，不能提供良好的拟合，说明还有其他因素在影响价格。

f. Provide an interpretation for the slope of the estimated regression equation.

回归方程的斜率是-0.05877，说明 miles 没增加 1，price 就相应的减少 0.05877。

g. Suppose that you are considering purchasing a previously owned 2007 Camry that has been driven 60,000 miles. Using the estimated regression equation developed in part (c), predict the price for this car. Is this the price you would offer the seller.

Price=16.46976-0.05877*(60000/1000)=12.94356, 预估的价格为：12.94356*1000=12943.56 从 a 中的线型图可以看出，价格和里程并不是线性的，而是曲线的，说明影响价格的因素比较复杂，还存在其他因素会影响价格，因此预估的价格不是最终的成交价，可以最为参考价。

Question #9:

附件 WE.xlsx 是某提供网站服务的 Internet 服务商的客户数据。数据包含了 6347 名客户在 11 个指标上的表现。其中” 流失 “指标中 0 表示流失，”1“表示不流失，其他指标含义看变量命名。

```
### 导入数据，并对变量进行简短统计
we<- read_xlsx("D:/xuying/data/WE.xlsx")
summary(we)
```

```
#>      客户ID           流失          当月客户幸福指数 客户幸福指数相比上月变化
#>  Min.   :   1    Min.   :0.00000   Min.   :  0.00   Min.   :-125.000
#>  1st Qu.:1588    1st Qu.:0.00000   1st Qu.: 24.50   1st Qu.:  -8.000
#>  Median :3174    Median :0.00000   Median : 87.00   Median :   0.000
#>  Mean   :3174    Mean   :0.05089   Mean   : 87.32   Mean   :   5.059
#>  3rd Qu.:4760    3rd Qu.:0.00000   3rd Qu.:139.00   3rd Qu.:  15.000
#>  Max.   :6347    Max.   :1.00000   Max.   :298.00   Max.   : 208.000
#>   当月客户支持      客户支持相比上月的变化 当月服务优先级
#>  Min.   : 0.0000   Min.   :-29.000000     Min.   :0.0000
#>  1st Qu.: 0.0000   1st Qu.:  0.000000     1st Qu.:0.0000
#>  Median : 0.0000   Median :  0.000000     Median :0.0000
#>  Mean   : 0.7063   Mean   : -0.006932     Mean   :0.8128
#>  3rd Qu.: 1.0000   3rd Qu.:  0.000000     3rd Qu.:2.6667
#>  Max.   :32.0000   Max.   : 31.000000     Max.   :4.0000
#>  服务优先级相比上月的变化  当月登录次数     博客数相比上月的变化
#>  Min.   :-4.00000          Min.   :-293.00   Min.   :-75.0000
#>  1st Qu.: 0.00000          1st Qu.:  -1.00   1st Qu.:  0.0000
#>  Median : 0.00000          Median :   2.00   Median :  0.0000
#>  Mean   : 0.03017          Mean   :  15.73   Mean   :  0.1572
#>  3rd Qu.: 0.00000          3rd Qu.:  23.00   3rd Qu.:  0.0000
#>  Max.   : 4.00000          Max.   : 865.00   Max.   :217.0000
#>  访问次数相比上月的增加  客户使用期限    访问间隔变化
#>  Min.   :-28322.00      Min.   : 5.0   Min.   :-646.000
#>  1st Qu.:   -11.00      1st Qu.:10.0   1st Qu.:   2.000
#>  Median :     0.00      Median :16.0   Median :   2.000
#>  Mean   :    96.31      Mean   :18.9   Mean   :   3.765
#>  3rd Qu.:    27.00      3rd Qu.:25.0   3rd Qu.:   5.000
#>  Max.   :230414.00      Max.   :72.0   Max.   :  63.000
```

```
skimr::skim(we)
```

表 5: Data summary

Name	we
Number of rows	6347
Number of columns	13
Column type frequency:	
numeric	13
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
客户 ID	0	1	3174.00	1832.37	1	1587.5	3174	4760.50	6347	
流失	0	1	0.05	0.22	0	0.0	0	0.00	1	
当月客户幸福指数	0	1	87.32	66.28	0	24.5	87	139.00	298	
客户幸福指数相比上月变化	0	1	5.06	30.83	-125	-8.0	0	15.00	208	
当月客户支持	0	1	0.71	1.72	0	0.0	0	1.00	32	
客户支持相比上月的变化	0	1	-0.01	1.87	-29	0.0	0	0.00	31	
当月服务优先级	0	1	0.81	1.32	0	0.0	0	2.67	4	
服务优先级相比上月的变化	0	1	0.03	1.46	-4	0.0	0	0.00	4	
当月登录次数	0	1	15.73	42.12	-293	-1.0	2	23.00	865	
博客数相比上月的变化	0	1	0.16	4.66	-75	0.0	0	0.00	217	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
访问次数相比上月的增加	0	1	96.31	3152.41	-28322	-11.0	0	27.00	230414	
客户使用期限	0	1	18.90	11.16	5	10.0	16	25.00	72	
访问间隔变化	0	1	3.76	17.97	-646	2.0	2	5.00	63	

a. 通过可视化探索流失客户与非流失客户的行为特点（或特点对比），你能发现流失与非流失客户行为在哪些指标有可能存在显著不同？

```
d9 <- we %>%
  rename(id="客户ID",
         churn="流失",
         happy_index="当月客户幸福指数",
         chg_hi="客户幸福指数相比上月变化",
         support="当月客户支持",
         chg_supprt="客户支持相比上月的变化",
         priority="当月服务优先级",
         chg_priority="服务优先级相比上月的变化",
         log_in_fre="当月登录次数",
         chg_blog_fre="博客数相比上月的变化",
         chg_vis="访问次数相比上月的增加",
         y_age="客户使用期限",
         chg_interval="访问间隔变化"
         )

### stats <- d9 %>%
###   group_by(churn) %>%
###   summarise(
###     Mean_happy_index = mean(happy_index),
###     Mean_chg_hi = mean(chg_hi),
###     Mean_support = mean(support),
###     Mean_chg_supprt = mean(chg_supprt),
###     Mean_priority = mean(priority),
###     Mean_chg_priority = mean(chg_priority),
###     Mean_log_in_fre = mean(log_in_fre),
###     Mean_chg_blog_fre = mean(chg_blog_fre),
###     Mean_chg_vis = mean(chg_vis),
###     Mean_y_age = mean(y_age),
###     Mean_chg_interval = mean(chg_interval)
###   )


summary_d9 <- d9 %>%
  group_by(churn) %>%
  summarise(
    across(
      c(happy_index, chg_hi, support, chg_supprt, priority, chg_priority,log_in_fre, chg_blog_fre,
      mean,
      na.rm=TRUE
    )
  )
round(summary_d9,2)
```

```
#> # A tibble: 2 x 12
#>   churn happy_index chg_hi support chg_supprt priority chg_priority log_in_fre
#>   <dbl>       <dbl>  <dbl>   <dbl>      <dbl>    <dbl>        <dbl>      <dbl>
#> 1     0        88.6   5.53    0.72      -0.01     0.83         0.03      16.1 
#> 2     1        63.3  -3.74    0.37       0.04     0.5         -0.02       8.06
#> # i 4 more variables: chg_blog_fre <dbl>, chg_vis <dbl>, y_age <dbl>,
#> #   chg_interval <dbl>
```

b. 通过均值比较的方式验证上述不同是否显著。

```
continuous_vars <- colnames(d9[3:13])

# 使用 lapply 函数进行批量 t 检验
t_test_results <- lapply(continuous_vars, function(var) {
  # 执行 Welch t 检验
  test_result <- t.test(as.formula(paste(var, "~ churn")), data = d9, var.equal = FALSE)
    # 提取并整理检验结果
  result_df <- data.frame(
                          Variable = var,
                          Statistic = round(test_result$statistic,2),
                          P.Value = round(test_result$p.value,2),
                          Conf.Int.Lower = round(test_result$conf.int[1],2),
                          Conf.Int.Upper = round(test_result$conf.int[2],2),
    stringsAsFactors = FALSE
  )
    return(result_df)
})

# 将所有检验结果合并为一个数据框
t_test_results_combined <- do.call(rbind, t_test_results)
t_test_results_combined
```

```
#>         Variable Statistic P.Value Conf.Int.Lower Conf.Int.Upper
#> t    happy_index      7.62    0.00          18.80          31.87
#> t1        chg_hi      5.78    0.00           6.12          12.42
#> t2       support      5.51    0.00           0.23           0.48
#> t3    chg_supprt     -0.63    0.53          -0.19           0.10
#> t4      priority      5.14    0.00           0.20           0.46
#> t5  chg_priority      0.64    0.52          -0.10           0.20
#> t6    log_in_fre      3.57    0.00           3.63          12.53
#> t7  chg_blog_fre      2.53    0.01           0.06           0.49
#> t8       chg_vis      1.91    0.06          -5.46         410.22
#> t9         y_age     -2.98    0.00          -2.55          -0.52
#> t10 chg_interval     -4.10    0.00          -7.36          -2.59
```

除了 chg_supprt 和 chg_priority 的 p_value 大于 0.05，其他指标的 p_value 均小于 0.05，其他指标都是显著的

c. 以”流失“为因变量，其他你认为重要的变量为自变量（提示：a、b 两步的发现），建立回归方程对是否流失进行预测。

```
# 将因变量 churn 转换为因子类型
d9$churn <- as.factor(d9$churn)
# 建立逻辑回归模型，选择您认为重要的自变量
d9_model <- glm(churn ~ happy_index + chg_hi + support + priority + log_in_fre + chg_blog_fre + ch

# 查看模型摘要
summary(d9_model)
```

```
#>
```

```
#> Call:
#> glm(formula = churn ~ happy_index + chg_hi + support + priority + 
#>     log_in_fre + chg_blog_fre + chg_vis + y_age + chg_interval, 
#>     family = binomial, data = d9)
#> 
#> Coefficients:
#>                Estimate Std. Error z value Pr(>|z|)    
#> (Intercept)  -2.874e+00  1.215e-01 -23.661  < 2e-16 ***
#> happy_index  -5.225e-03  1.161e-03  -4.500 6.78e-06 ***
#> chg_hi       -9.501e-03  2.424e-03  -3.920 8.87e-05 ***
#> support      -3.522e-02  7.438e-02  -0.474  0.63581    
#> priority     -3.727e-02  7.514e-02  -0.496  0.61985    
#> log_in_fre    9.104e-04  1.952e-03   0.466  0.64098    
#> chg_blog_fre -2.357e-05  2.080e-02  -0.001  0.99910    
#> chg_vis      -1.170e-04  4.069e-05  -2.877  0.00401 ** 
#> y_age         1.418e-02  5.260e-03   2.696  0.00701 ** 
#> chg_interval  1.700e-02  4.277e-03   3.975 7.03e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> (Dispersion parameter for binomial family taken to be 1)
#> 
#>     Null deviance: 2553.1  on 6346  degrees of freedom
#> Residual deviance: 2445.9  on 6337  degrees of freedom
#> AIC: 2465.9
#> 
#> Number of Fisher Scoring iterations: 6
```

d. 根据上一步预测的结果，对尚未流失（流失 =0）的客户进行流失可能性排序，并给出流失可能性最大的前 100 名用户 ID 列表。

```r
# 计算所有客户的流失概率
d9$predicted_probabilities <- predict(d9_model, newdata = d9, type = "response")

# 过滤尚未流失的客户（流失 =0）
not_churned_customers <- d9[d9$churn == 0, ]

# 根据流失概率进行排序
```

```r
sorted_customers <- not_churned_customers[order(-not_churned_customers$predicted_probabilities), ]

# 提取流失可能性最大的前 100 名用户 ID
top_100_customers <- head(sorted_customers$id, 100)

# 输出前 100 名用户 ID
print(top_100_customers)
```

```
#>  [1] 2287  109 1971 2025    1  929 2076   76   14   18    3 2244   21 1287 1929
#> [16] 1459   51  128  183   59   55  121 2240 1520 2599 1236  137 1862 2080 1143
#> [31]  154 1286 2546  146  119  171  190   42    5    2  123  101 1616   95 2680
#> [46] 2838   61 2289 1438 1392 2481 2924  106 3671  203 1393   69 1574 1204   68
#> [61] 2255 1395 1478 2235   89  798 1141 2739   62 4245 1151 2830 1693 3042   12
#> [76]  142 1908   10  868 2286 3076   57 2242 1951 3124 1019 1110 2062 2903 2913
#> [91] 2047  104 1953 2656 1155 2744 1446 2306  163  240
```